基于突发主题词和凝聚式层次聚类的微博突发事件检测研究*

丁晟春　龚思兰　李红梅
(南京理工大学经济管理学院　南京 210094)

摘要:【目的】实时、准确、高效地检测出海量微博中的突发事件, 为舆情应急管理提供重要的决策信息支持。【方法】引入参照时间窗机制, 设计词频、文档频率、话题标签(Hashtag)、词频增长率 4 类特征的选择与计算方法, 基于动态阈值实现对突发主题词的抽取。在此基础上, 将微博文本表示为突发主题词的特征向量, 使用凝聚式层次聚类算法实现了突发事件的检测。【结果】将实验结果结合实例进行分析, 突发事件检测达到 80%的准确率, 验证该方法的可行性和有效性。【局限】由于语料数据和研究范围的限制, 还未实现对所检测突发事件的自动描述, 对网民情感、事件间语义关系等要素的分析及考量也存在一定欠缺。【结论】本研究突破以往相关研究中文本内容质量、文本形式、突发特征抽取结果的局限, 提升微博突发事件检测的效率。

关键词: 突发事件检测　突发主题词　凝聚式层次聚类　网络舆情　微博
分类号: G202

1 引 言

微博作为新型社交媒体平台具有使用方式便捷、传播迅速、交流互动性强、内容全面等特征, 已经成为突发事件信息快速聚集和传播的重要渠道。突发事件是指突然发生, 造成或者可能造成严重社会危害, 需要采取应急处置措施予以应对的自然灾害、事故灾难、公共卫生事件和社会安全事件。突发事件的产生具有瞬间性, 发生中的爆发点具有偶然性, 发生的时间、地点等非常突然。当有突发事件发生时, 广大网民越来越习惯通过微博实时发布和获取突发事件相关信息, 并针对突发事件发表个人观点态度。此外, 突发事件的频繁爆发使得针对微博的网络舆情舆论分析获得了各界的密切关注。在突发事件爆发的第一时间从海量微博中准确而高效地检测出突发事件, 不仅可以帮助用户实时获取重要的突发事件资讯, 消除突发事件带来的恐慌心理, 还能够协助应急管理机构实时把握突发事件的发展态势, 合理地控制和引导舆论发展动向, 为舆情应急管理提供决策信息支持, 这些对于发挥网络舆情在保证民众知情权利和维护社会稳定健康发展等方面的积极作用具有重要意义。

2 相关研究

针对微博的突发事件检测研究已经取得一定的成果, 主要分为以文档为中心[1]和以特征为中心的检测研究[2]。

以文档为中心的突发事件检测技术是直接对文档进行聚类, 将类簇看作突发事件, 在此基础上抽取事件特征, 用以表示检测出的突发事件[3-4]。Petrović 等[5]提出一种基于 LSH(Locality-Sensitive Hashing) 的

通讯作者: 丁晟春, ORCID: 0000-0002-4269-021X, E-mail: todingding@163.com。

*本文系国家社会科学基金项目“基于社会网络分析的网络舆情主题发现研究”(项目编号:15BTQ063)、中央高校基本科研业务费专项资金资助项目“大数据时代基于深度融合的创新型知识服务体系及其运行机制研究”(项目编号:30916011330)和国家社会科学基金重点项目“大数据环境下社会舆情与决策支持方法体系研究”(项目编号:14AZD084)的研究成果之一。

Twitter 文本聚类算法，该方法优化了社交媒体上的事件检测时间效率，并能保持算法的时间和空间复杂度恒定。Phuvipadawat 等[6]研究了 Twitter 上突发类新闻事件无监督聚类方式，选择了一般特征和微博特有特征，使用 TF-IDF 方法赋予每种特征词对应权重，取得了不错的聚类结果。葛高飞[7]提出了改进的 TC-LDA 算法，解决突发事件检测中的噪音问题。

以特征为中心的突发事件检测技术[8]重点在于检测突发事件在实时数据流上随时间变化的突发特征，即抽取突发主题词[9]，通过对突发主题词聚类或者使用突发主题词对文本进行表示后再利用聚类算法，达到突发事件检测目的。该方法能够避免数据稀疏问题，但微博文本短小而数据量大，含很多广告、网络欺诈等大量噪音，更具实时性，所以微博上的突发事件检测更易受到垃圾信息的影响[10]。针对噪声数据，研究者注重利用时序信息，结合微博自带的属性功能 Hashtag 等[11]挖掘事件发生时期所呈现的突发特征。Kleinberg[12]很早就发现文档流会出现突然持续一段时间后消失的特征，提出了经典的 Bursty 挖掘方法。He 等[13]分析词语在时间序列上的趋势，应用于无监督的突发事件识别算法中。Mathioudakis 等[14]实现的“Twitter Monitor”系统利用在特定时间内 Twitter 中出现的频次异常高的突发词进行聚类，实时发现新兴突发事件。Long 等[15]在事件检测上引入词语的文档频率、Hashtag、信息熵因素以提取代表突发事件的主题词，构建词共现图应用聚类算法获得了微博中的事件。赵文清等[16]使用相对词频和词频增长率抽取突发事件的主题词，基于词语间的共现图聚类，将类簇看作微博新闻事件。Yao 等[17]通过监测用户产生的信息标记的 Hashtag 词变化检测微博中发生的事件。王勇等[18]从词频统计、词频增长率和 TF-PDF 三方面计算词语的权重用于抽取突发词集，提出“绝对聚类”算法以较准确地检测突发事件。郭跚秀等[19]将微博文本特征、微博传播特征和用户影响力融合计算，用于抽取突发词，使用凝聚式层次聚类对突发词聚类来检测微博的突发事件。

综上，现有针对突发事件检测的研究仍存在一定的局限性，多受制于文本内容质量、文本形式、突发特征抽取结果等因素的影响。基于此，研究引入微博事件三要素过滤策略实现对微博文本内容质量的把控；同时结合微博文本形式特点，通过繁简字体转换、分词、停用词处理和词性过滤等方式对微博数据进行预处理，过滤可能影响突发特征的噪声信息，统一文本形式；接着在综合考虑词语主题表达能力和突发性的基础上，引入参照时间窗机制，设计了词频、文档频率、话题标签(Hashtag)、词频增长率 4 类特征的选择与计算方法，基于动态阈值提取有效表征事件的突发主题词作为突发特征；最后将微博文本表示为特征向量，构造微博文本相似度矩阵，使用凝聚式层次聚类算法实现对微博突发事件的检测研究。

3　突发事件检测研究框架

突发事件的舆情是互联网用户围绕特定突发事件而持有各自的观点，进行相互交流与沟通产生的结果，并形成一定的信息流，表现出周期性特征。突发事件在微博平台上爆发后，一些能够用于描述事件的特征便被广泛提起，从语言学角度分析，这些特征是特定时间内突发事件表现在微博文本内容上的突发类词语，但单纯使用词语并不能将事件区分开，需要利用突发类词语定位到对应时间段内的微博文本，利用该文本聚类实现对突发事件的检测。

具体检测框架如图 1 所示。研究主要解决以下几个方面的问题:

(1) 微博数据垃圾信息多，突发特征容易受到噪声的影响，因此抽取突发主题词之前需要特别注意微博数据的噪声、垃圾信息过滤，并进行文本分词、词性标注等预处理操作。

(2) 针对微博突发事件的传播特征，对采集的微博进行时间窗划分，构造时间序列上的微博数据流，通过捕获不同时间窗词语的时序分布及突发规律，利用词语的词频、文档频率、话题标签、词频增长率 4 类特征，基于动态阈值抽取事件的突发主题词。

(3) 对描述突发事件的文本进行过滤后，利用突发主题词作为突发特征表示微博文本，使用凝聚式层次聚类策略将微博文本聚成类簇，将聚类结果作为突发事件。

(4) 通过实验对研究方法进行验证，并结合实例对突发事件检测效果进行分析。

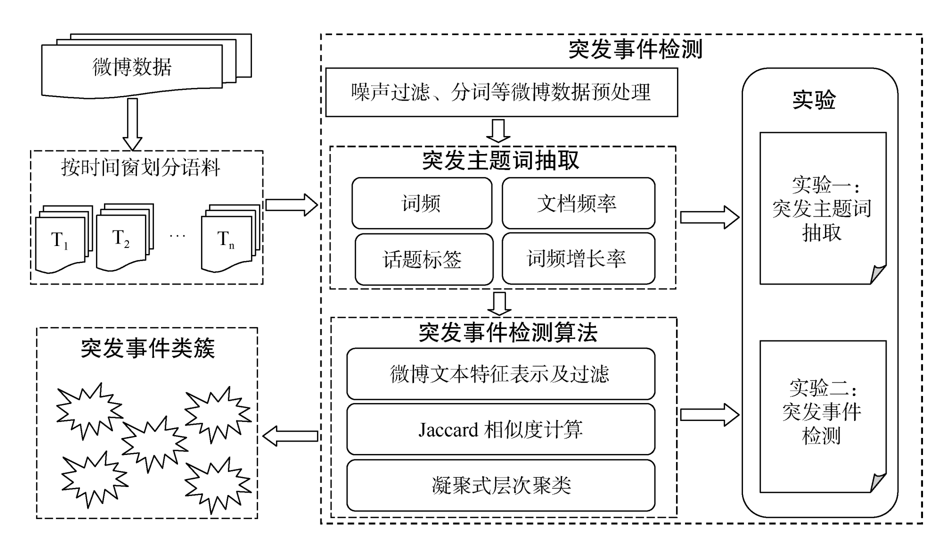


图 1 微博平台上的突发事件检测框架

4 基于突发特征的突发事件检测方法

4.1 突发主题词抽取

(1) 突发主题词特征分析

突发主题词是在某个时间窗内被大量使用，而在其之前的时间窗内很少被使用的实词[9]。基于微博本身具有及时性和裂变传播的特征，在抽取突发事件的突发主题词之前，需要先将连续的微博数据流按照独立的时间段划分。本文将微博数据按照时间信息划分成 m×t 个时间窗，其中 m 以“天”为单位，为了更加细粒度地实时检测微博中事件的发生时间，t 则可以在“天”的基础上根据所需划分为更细时间段的个数，以“天”、“小时”、“分钟”或“秒”为单位。为了使从微博中提取的词语更全面地描述突发事件，本文从词频、文档频率、话题标签、词频增长率 4 个方面制定突发主题词衡量标准，判断文本中一个词语能否成为突发主题词。

①词 频

词频能够衡量一个词汇在文档中的重要程度。在统计意义下，某一词汇若频繁出现，在某种程度上就意味着这个词语与文本所表达主题相关的可能性较高，因此本文采用词频作为突发主题词选择的衡量方法之一。

②文档频率

对于突发事件来讲，若包含某一词语的微博数量在当前时间窗内比较多，说明这个词语越可能是某一突发事件的特征词。为了保证所选特征词的主题表现力，本文对文档频率进行调整，引入熵的概念，衡量一个词语对于该段时间窗的突发事件主题的表现力，熵越大说明这个词语越能表达突发事件的主题信息。

③话题标签

作为微博最具有特色的功能属性之一，话题标签(Hashtag)能够让用户为所发布的信息内容创建一个主题标签[20]，与事件关系越紧密的特征词越容易出现在微博话题标签中，本文充分利用微博话题标签特征，通过计算词语的话题标签权重衡量该词语与某个突发事件相关的程度[21]，其计算公式如下：

$$HT_{ij}=\begin{cases}\dfrac{h_i}{N} & l(w_i)=1\\[2mm] \dfrac{h_i'}{N+1} & l(w_i)=0\end{cases} \tag{1}$$

其中，HT_{ij} 是时间窗 j 上词语 w_i 的话题标签权重，$l(w_i)$是判别函数，$l(w_i)=1$ 表示至少存在一个话题标签中包含词语 w_i，$l(w_i)=0$ 表示话题标签中均不包含词语 w_i，h_i 是词语 w_i 出现在话题标签中的次数计数，N 为在当前时间窗 j 时间段内话题标签的总数，h_i' 是当前时间窗内包含词语 w_i 且含有话题标签#的微博条数。

④词频增长率

词语的突发性会随着时间变化而呈现急剧增加的状态，其最明显的特征就是利用词频增量来筛选当前时间窗内的突发主题词，词频增量通常使用该词在相邻时间窗的词频比例变化来计算[18]。同时，为了避免事件持续期内相邻时间窗对词频增长率结果的干扰，本文将参照时间窗机制中的相对时间窗和相邻时间窗进行结合对比，计算公式如下：

$$FT_{ij}=\lambda_1\frac{f_{ij}}{1+f_{i(j-1)}}+\lambda_2\frac{f_{ij}}{1+f_{i'j'}} \tag{2}$$

其中，FT_{ij} 表示词语 w_i 在当前时间窗 j 的词频增长率，f_{ij} 是词语 w_i 在时间窗 j 内出现的词频，$f_{i(j-1)}$是词语 w_i 在前一个时间窗 j–1 上的词频，若以天为时间单位，则 $f_{i'j'}$是词语 w_i 在

j–2 时间窗上的词频，若以小时为时间单位，则 $f_{ij'}$对应前一天的第 j′个时间窗的词频。λ_1 和 λ_2 分别是调节系数，$\lambda_1+\lambda_2=1$。

根据上述分析可以发现微博中词频、文档频率、话题标签权重、词频增长率均较高的词语更有可能成为描述事件的突发主题词。词语的突发主题度将综合这 4 种特征的归一化结果进行计算，计算公式如下：

$$BTword_{ij} = F'_{ij} + DF'_{ij} + HT'_{ij} + FT'_{ij} \quad (3)$$

其中，$BTword_{ij}$ 表示词语 w_i 的在时间窗 j 的突发主题度，F'_{ij}、DF'_{ij}、HT'_{ij}、FT'_{ij} 分别是归一化后的词频、文档频率、话题标签权重、词频增长率。将最终得到的突发主题词集合 BTword 表示为：BTword={$word_1$, $word_2$, $word_3$, …$word_k$}，其中 $word_k$ 表示在当前时间窗 j 内第 k 个突发主题词。

(2) 突发主题词抽取算法

一个词语能否成为突发主题词需要先满足设定的阈值 δ 标准，然后再计算这些满足标准的所有词语的突发主题度。阈值 δ 包括：当前时间窗内所有词的词频的平均值 δ_1；当前时间窗内所有词的文档频率的平均值 δ_2；调节词语的突发特性的经验动态阈值 δ_3；当前时间窗内满足前 3 个阈值的词语突发主题度的平均值 δ_4。突发主题词抽取算法具体流程如下：

①输入微博数据流，按正文发布时间划入不同的时间窗口，然后对每一窗口中的微博预处理后进行统计，得到每个时间窗口内的总词语列表 W。

②读取词语序列 W 中的词 w_i，执行步骤③。

③计算词语 w_i 的词频，判断是否大于阈值 δ_1，若大于则保留该词语，执行步骤④，否则过滤该词语 w_i，设置 i=i+1，跳至步骤②。

④计算词 w_i 的文档频率，判断是否大于阈值 δ_2，若大于则保留该词语，执行步骤⑤，否则过滤该词语 w_i，设置 i=i+1，跳至步骤②。

⑤计算词 w_i 的词频增长率，判断是否大于阈值 δ_3，若大于则保留该词语，执行步骤⑥，否则过滤该词语 w_i，设置 i=i+1，跳至步骤②。

⑥对满足以上阈值标准的词语 w_i 先计算话题标签率，然后综合计算其突发主题度，判断是否大于阈值 δ_4，若大于则保留该词语，执行步骤⑦，否则过滤该词语 w_i，设置 i=i+1，跳至步骤②。

⑦将该词语 w_i 添加到事件突发主题词列表 BTword 中，最终输出该时间窗所有的突发主题词。

按照以上流程处理该时间窗内的所有词语，保留满足阈值的全部词语作为突发主题词，这些突发主题词既有较高的主题表现力，又能体现事件的突发特性，因此能较好地表征突发事件。

4.2　突发事件检测

(1) 基于突发主题词的微博文本特征表示

对于某个时间窗中的任意一条微博文本，基于当前时间窗中的突发主题词集 BTword={$word_1$, $word_2$, $word_3$, …$word_k$}进行文本特征向量构建，定义微博文本的向量形式化表示如下：

$$text_i = \{term_{i1}, term_{i2}, term_{i3}, \cdots, term_{ik}\} \quad (4)$$

其中，$text_i$ 表示第 i 个微博文本，$term_{ik}$ 表示第 i 个微博文本中是否包含第 k 个突发主题词的情况，$term_{ik}=1$ 则表示包含该突发主题词，$term_{ik}=0$ 则表示不包含该突发主题词。例如，如果时间窗 j 中的突发主题词集为{护士，南京，袁亚平，官员，瘫痪}，微博文本 $text_i$ 中包含的突发主题词有{护士，南京，瘫痪}，那么 $text_i$ 可以表示为：$text_i$={1, 1, 0, 0, 1}。

借鉴文献[18]中的微博文本过滤原则，认为描述事件的一条微博文本应至少包含“5W1H”中的任意 3 个要素，但要素类型在微博文本特征向量中不再区分，具体到微博文本的突发主题词特征向量时，应该至少包含 3 个突发主题词。通过剔除所有微博文本中包含突发主题词个数小于 3 的微博，降低微博文本–突发主题词矩阵的稀疏性，可以有效提高突发事件检测的效率，同时也保证了检测结果的完整性。

(2) 基于凝聚式层次聚类的突发事件检测算法

在对微博文本进行特征表示后发现，微博中用户表达突发事件的语言比较相近，表现出一种“围观”现象，关联事件的微博一般会集中出现，微博文本中的词语也通常围绕某些事件特征词语，重复度很高。因此，本文认为不同微博文本中包含相同突发主题词的个数越多，它们越可能描述的是同一个突发事件。

在相似度计算方法选择方面，采用 Jaccard 系数方法来判断微博文本特征向量间的相似度，更符合突发事件聚合的真实情况，能够反映出微博文本间真实的相似性，即讨论相同事件的两个微博文本重合度应该是较高的，具体 Jaccard 相似度计算公式如下：

$$S(text_i, text_j) = \frac{|text_i \cap text_j|}{|text_i \cup text_j|} \quad (5)$$

其中，$S(text_i, text_j)$是两个微博文本之间的相似度，

$text_i$表示微博文本特征向量, $text_i \cap text_j$表示$text_i$, $text_j$的交集, $text_i \cup text_j$表示$text_i$, $text_j$的并集。

基于凝聚式层次聚类的事件检测算法流程如下:

①输入时间窗j的微博集,对微博文本进行突发主题词特征向量表示,记为$text_i$,将突发主题词数少于3个的$text_i$向量进行过滤,形成微博文本–突发主题词矩阵D。

②初始化每个微博文本特征向量作为一个类,利用Jaccard系数计算两两微博文本特征向量的相似度值$S_{i,j}$,构建微博文本的相似度矩阵S。

③查找相似度矩阵S中的最大值$\max\{S_{i,j}\}$。

④依据层次聚类的合并规则,将事件类i、事件类j合并成新的向量,同时重新计算该新向量与已有事件类向量的相似度,重新调整相似度矩阵S。

⑤根据矩阵S中列数或行数判断是否满足预设阈值,若满足,执行步骤⑥,否则跳转至步骤③。

⑥通过该聚类过程,最终将所有的微博文本聚成n个类簇,将微博文本特征向量$text_i$映射为原始的微博文本,输出最终的聚类结果,其中每个类簇代表一个突发事件。

5 实验及结果分析

5.1 实验数据源及结果评价指标

(1) 数据源及预处理

在针对微博数据的突发事件检测研究领域,尚无国际公认的标准测试语料。本文的实验数据来源于新浪微博,基于其开放平台API接口的微博爬虫实现数据爬取,受限于API接口的频次及数量访问限制,只获取了新浪微博的部分数据(2014年2月25日至3月11日期间的180多万条微博数据)。从实验的角度出发,这些微博数据可以作为全部数据的一个样本代表,用于支撑本文的实验分析与研究。

微博数据中充斥着大量的垃圾和噪声信息,会对突发事件检测结果造成严重的影响,在对微博进行突发事件检测之前,需要对微博数据进行预处理,预处理操作包括噪声过滤、繁简字体转换、分词及停用词处理和词性过滤等。微博中的噪声主要有@XXX噪声、URL链接噪声和表情符号,本文针对特定的噪声类型设置正则表达式对其进行过滤;再进一步根据《通用规范汉字表》[22],提取所有的繁体字和其对应的简体字,然后将所有的繁体字和简体字进行对应,分别构建《繁体中文字表》和《简体繁体中文字对应表》实现微博繁简字体转换;接着,利用NLPIR汉语分词系统[23]对微博进行分词处理并根据其标注的词性实现词性过滤,保留名词和动词;最后使用停用词词表,利用词汇匹配方式过滤停用词,完成对微博数据的预处理。

(2) 实验结果评价指标

①突发主题词抽取评价

传统的评价指标包括准确率(Precision)、召回率(Recall)以及F值(F-measure)三个参数,由于当前时间窗内无法获取所有的突发主题词,因此突发主题词抽取的评价指标中,召回率是难以直接计算的,所以本文利用未进行突发主题度平均值阈值δ_4,过滤当前情况下抽取正确的突发主题词作为该时间窗的总体突发主题词,以此计算召回率,并认为符合阈值δ_4情况下抽取到的突发主题词为实验最终需要的词语,采用准确率、召回率、F值进行评价,突发主题词语是否抽取正确则利用人工判断该时间窗上抽取的突发主题词能否描述或概括现实生活中发生的突发事件。具体的评价公式计算如下:

$$\text{Precision(BTword)} = \frac{k}{K} \tag{6}$$

$$\text{Recall(BTword)} = \frac{k}{S} \tag{7}$$

$$\text{F-measure(BTword)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

其中,Precision(BTword)表示突发主题词抽取准确率,Recall(BTword)表示突发主题词抽取召回率,k表示当前时间窗抽取正确的突发主题词个数,K是当前时间窗抽取的突发主题词总数,S是未进行突发主题度平均值阈值δ_4过滤的词语中所有抽取正确的突发主题词总数。

②突发事件检测评价

突发事件检测结果方面,因为现实生活中发生的突发事件是无法事先预知的,即某一时间窗口内微博中所有的突发事件数量是难以事先获取到的,对于该结果的召回率也无法直接计算得到,因此本实验只选用准确率来评估检测的突发事件正确与否。将实验检测的突发事件结果进行人工比较,通过判断检测的突发事件是否反映了真实的突发事件,若是则视为识别正确,否则视为错误。突发事件检测评价公式计算如下:

$$\text{Precision(event)} = \frac{e}{E} \tag{9}$$

其中,Precision(event)是突发事件检测的准确率,e表示当前时间窗正确检测的突发事件个数,E是当前时间窗检测出的突发事件个数。

5.2 突发主题词抽取实验及结果分析

针对微博文本中事件的突发主题词抽取,本文选用词频、文档频率、话题标签率、词频增长率4类特征,如表1所示。设计5组特征组合计算方法进行对比实验,由此说明本文所选取的特征计算方法是有效

的。方法 1 用来考察 4 个特征计算组合方法对突发主题词抽取的影响，方法 2 至方法 5 用来分析词语的词频增长率、话题标签率、文档频率、词频特征计算方法对突发主题词抽取的影响。

表 1 突发主题词的特征计算方法对比组合设计

方法编号	特征计算方法组合
1	词频，文档频率，话题标签率，词频增长率
2	词频，文档频率，话题标签率
3	词频，文档频率，词频增长率
4	词频，话题标签率，词频增长率
5	文档频率，话题标签率，词频增长率

以天为时间窗，使用 2014 年 2 月 25 日至 2 月 27 日的数据进行实验，设置词频增长率计算公式中调节系数 $\lambda_1=0.5$，$\lambda_2=0.5$，词频增长率阈值 $\delta_3=0.5$。在不同的特征计算方法下针对 2 月 27 日数据进行突发主题词抽取的统计结果如图 2 所示：

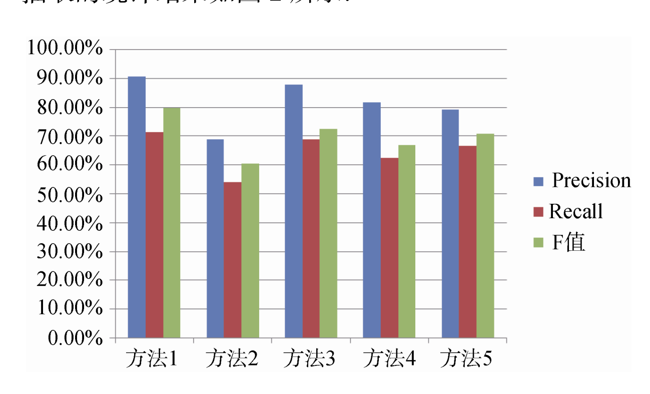


图 2 不同特征计算方法下突发主题词抽取结果

根据图 2 可以发现，方法 2 与方法 1 对比，准确率、召回率都下降了很多，这说明词频增长率计算方法对于抽取突发主题词特别重要，利用词频增长率判断词语的突发特性，可以提高突发主题词抽取的准召率。方法 3 至方法 5 的结果说明了本方法中话题标签率、文档频率、词频三种特征计算方法的有效性，对比准确率发现词频对提高准确率的作用最大，其次是文档频率、话题标签率。从召回率看出，文档频率能够提高一个词的重要性，更能将其选择为突发主题词。综合来看，显然方法 1 组合使用这 4 种特征计算方法所达到的实验效果最好。

5.3 突发事件检测实验及结果对比分析

该实验验证突发事件检测算法的可行性，通过划分不同的时间窗，使用 4 个特征的组合计算方法对时间窗内的突发主题词进行抽取。本部分实验选取凝聚式层次算法和 K-means 算法进行聚类和对比，以天为时间窗，选取 2014 年 2 月 27 日的数据为例，分别设置聚类的类簇为 5、10、15、20 进行实验，利用准确率评价基于聚类方法的突发事件检测效果，最终的统计结果如图 3 所示：

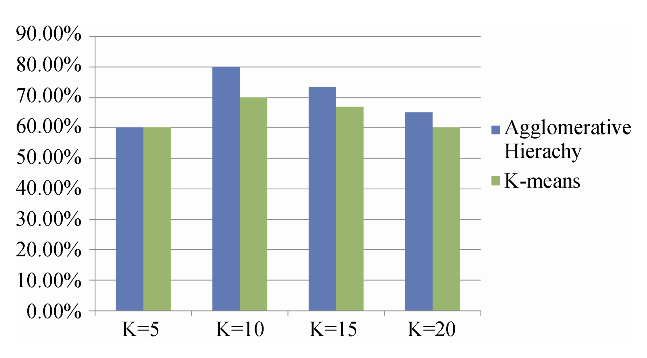


图 3 凝聚式层次聚类和 K-means 聚类算法下突发事件检测的准确率结果对比

根据实验结果可知，在突发事件检测算法上，凝聚式层次聚类的结果要优于 K-means，当聚类的类簇个数 K 取值为 10 的时候，突发事件检测的准确率结果达到 80%，获得一个较优的结果，随着 K 增大，准确率有所降低。对实验数据和结果进行分析发现，一个突发事件可能会涉及多个方面的内容，当 K 过度增大，会将一个突发事件划分成多个与事件相关但又无法成为一个独立事件的细粒度侧面信息，这些侧面信息的类簇有可能不是突发事件。

实验取得不错的效果，笔者认为主要有以下几个方面的原因：

(1) 针对突发特征容易受到微博中噪声数据的影响问题，研究了微博上的噪声处理方法，通过对繁体转简体操作，@XXX 符号、URL 链接及表情符号等噪声的过滤、词性过滤及停用词的处理，有效提高了微博文本的质量，为突发事件的检测提供了较好的数据基础。

(2) 充分利用突发特征的爆发规律，结合微博自身的 Hashtag 标签属性，提出了动态阈值的突发主题词抽取算法。选取词频、文档频率、话题标签和词频增长率特征计算方法，在此基础上设计抽取算法能够有效地从大量的词语中筛选既具有主题表现力，又具

有突发特性的高质量突发主题词，而且动态调节词频增长率阈值能够获得不同的突发主题词个数，进而影响检测的突发事件个数。

(3) 以突发主题词为基础，对微博文本进行特征向量表示，并结合设置的过滤策略保留有效的微博向量，利用Jaccard相似系数计算微博特征向量间的相似度，更好地体现了突发事件文本间的相似情况，在此基础上，利用凝聚式层次聚类实现了突发事件检测，确保了突发事件检测的可行性和有效性。

6 结　论

本文以微博为研究平台，针对突发事件检测进行研究，结合微博自身特征设计并实现了以突发特征为中心的突发事件检测方法，进行了突发主题词抽取和突发事件检测实验，获得了较高的准确率。由于语料数据和研究范围的限制，还未实现对所检测突发事件的自动描述，在对网民情感、事件间的语义关系等对突发事件检测有重要意义的要素分析及考量还有一定欠缺。因此，未来将进一步尝试结合网民的用户情感特征、事件间的语义关系辅助微博平台上的突发事件检测研究，以期获得更好的检测结果。

参考文献：

[1] Wang X, Zhai C X, Hu X, et al. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams[C]. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 784-793.

[2] Du Y, Wu W, He Y, et al. Microblog Bursty Feature Detection Based on Dynamics Model [C]. In: Proceedings of 2012 International Conference on Systems and Informatics (ICSAI). IEEE, 2012: 2304-2308.

[3] Aggarwal C C, Zhai C X. A Survey of Text Clustering Algorithms [A]. //Mining Text Data [M]. Springer US, 2012: 77-128.

[4] Yang Y, Pierce T, Carbonell J. A Study of Retrospective and On-line Event Detection[C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998: 28-36.

[5] Petrović S, Osborne M, Lavrenko V. Streaming First Story Detection with Application to Twitter[C]. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 181-189.

[6] Phuvipadawat S, Murata T. Breaking News Detection and Tracking in Twitter [C]. In: Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE, 2010: 120-123.

[7] 葛高飞. 突发事件微博新话题检测与跟踪系统的设计与实现[D]. 北京: 北京邮电大学, 2014. (Ge Gaofei. Design and Implementation of New Topic Detection and Tracking of Microblog Based on Emergency [D]. Beijing: Beijing University of Posts and Telecomunications, 2014.)

[8] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter[C]. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. 2011: 438-441.

[9] Du Y, He Y, Tian Y, et al. Microblog Bursty Topic Detection Based on User Relationship [C]. In:Proceedings of 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2011: 260-263.

[10] Benevenuto F, Magno G, Rodrigues T, et al. Detecting Spammers on Twitter [C]. In: Proceedings of the Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS). 2010(6): 12-20.

[11] Weng J, Lee B S. Event Detection in Twitter [C]. In: Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain. 2011: 401-408.

[12] Kleinberg J. Bursty and Hierarchical Structure in Streams [J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.

[13] He Q, Chang K, Lim E P. Analyzing Feature Trajectories for Event Detection [C]. In:Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007: 207-214.

[14] Mathioudakis M, Koudas N. Twittermonitor: Trend Detection over the Twitter Stream[C]. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. ACM, 2010: 1155-1158.

[15] Long R, Wang H, Chen Y, et al. Towards Effective Event Detection, Tracking and Summarization on Microblog Data [A]. //Web-Age Information Management [M]. Springer Berlin Heidelberg, 2011: 652-663.

[16] 赵文清, 侯小可. 基于词共现图的中文微博新闻话题识别[J]. 智能系统学报, 2012, 7(5): 444-449. (Zhao Wenqing, Hou Xiaoke. News Topic Recognition of Chinese Microblog Based on Word Co-occurrence Graph [J]. CAAI Transactions on Intelligent Systems, 2012, 7(5): 444-449.)

[17] Yao J, Cui B, Huang Y, et al. Bursty Event Detection from Collaborative Tags [J]. World Wide Web, 2012, 15(2): 171-195.

[18] 王勇, 肖诗斌, 郭跚秀, 等. 中文微博突发事件检测研究[J]. 现代图书情报技术, 2013(2): 57-62. (Wang Yong, Xiao Shibin, Guo Yixiu, et al. Research on Chinese Micro-blog Bursty Topics Detection [J]. New Technology of Library and Information Service, 2013 (2): 57-62.)

[19] 郭跚秀, 吕学强, 李卓. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用, 2014, 34(2): 486-490, 505. (Guo Yixiu, Lv Xueqiang, Li Zhuo. Bursty Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering [J]. Journal of Computer Applications, 2014, 34(2): 486-490, 505.)

[20] Small T A. What the Hashtag? A Content Analysis of Canadian Politics on Twitter [J]. Information, Communication & Society, 2011, 14(6): 872-895.

[21] 张志瑛. 基于主题模型和社区发现的微博热点事件检测研究[D]. 重庆: 西南大学, 2014. (Zhang Zhiying. Research on Hot Event Detection in Micro-blog Based on Topic Model and Community Discovery [D]. Chongqing: Southwest University, 2014.)

[22] 国家语言文字工作委员会.《通用规范汉字表》[K]. 2013.08. http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm. (National Languages Committee. The Common Standard Chinese Characters Table [K]. 2013.08. http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm.)

[23] NLPIR 汉语分词系统[CP/OL]. http://ictclas.nlpir.org/ downloads. (NLPIR Chinese Word Segmentation System [CP/OL]. http://ictclas.nlpir.org/downloads.)

作者贡献声明:

丁晟春: 论文选题, 提出研究思路与方法, 论文修订;
龚思兰: 负责资料数据收集、论文起草和论文修订;
李红梅: 数据采集与处理分析, 论文起草,。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[2-4]由作者自存储, E-mail: njustgsl@163.com。

[1] 龚思兰, 李红梅. data_results.rar. 数据处理结果.
[2] 龚思兰, 李红梅. Weibo_crawler.py. 微博语料抓取程序.
[3] 龚思兰. weibo_0225-0311.sql. 原始微博语料.
[4] 李红梅. weibo_process.jar. 语料处理程序.

收稿日期: 2016-06-12
收修改稿日期: 2016-07-11

A New Method to Detect Bursty Events from Micro-blog Posts Based on Bursty Topic Words and Agglomerative Hierarchical Clustering Algorithm

Ding Shengchun Gong Silan Li Hongmei
(School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: [**Objective**] This paper proposes a new method to detect real time bursty events accurately and efficiently from massive micro-blog posts. It provides decision-making information to public opinion emergency management. [**Methods**] First, we introduced the reference time window mechanism, and then designed an algorithm to process the data of word frequency, document frequency, Hashtags, and word frequency growth rates. Second, used this dynamic threshold based algorithm to extract bursty words. Third, transformed micro-blog texts to feature vector of the bursty words. Finally, we detected the bursty events using agglomerative hierarchical clustering algorithm. [**Results**] The bursty events detection method reached 80% of accuracy rate compared with real world cases. Thus, the proposed method was feasible and effective. [**Limitations**] We could not describe the detected emergencies automatically due to the limits of data and size of the current study. More research is needed to analyze users' emotion and semantic relationships among the bursty events. [**Conclusions**] Our study fills the knowledge gaps left by previous research, and improves the efficiency of retrieving bursty events from micro-blog posts.

Keywords: Bursty events detection Bursty topic words Agglomerative hierarchical clustering algorithm Public opinion Micro-blog

EBSCO 信息服务助力全球研究人员研究一带一路多地区贸易计划

EBSCO 于近期推出一个权威的国际化的一带一路参考资源库，收集了来自 60 多个国家的期刊和出版物。这个一带一路参考资源库，帮助研究人员更好地理解一带一路沿线国家的文化和经济状况，发现新的贸易机会。

一带一路倡议是由中华人民共和国提出的贸易和经济增长战略。该倡议旨在通过海上丝绸之路的发展，进一步连接中国大陆与一带("新丝绸之路")沿线的西欧贸易伙伴。即将推出的海上改进策略包括新的货运基础设施和区域港口建设，以支撑海外航运举措。

EBSCO 的一带一路参考资源库提供了 5 300 多份全文期刊，包括许多难以找到的一带一路沿线国家的本地出版物。该资源库还包含近 65 种全文报纸和 270 多份报告和会议集。建设该资源库是 EBSCO 致力于全球学术研究的又一举措。通过提供高质量内容，一带一路参考资源库能为研究人员提供多国环境下的全局和局部视角。

该资源库涵盖了多学科的内容，文章来源范围广泛，包括:《建筑科学与工程杂志》(中国)、《测绘，建筑及物业杂志》(马来西亚)、《教育科学》(土耳其)、《理论和应用信息技术杂志》(巴基斯坦)、《生物医学化学》(俄罗斯)、《GSTF 数学、统计学及运筹学杂志》(新加坡)、《中国经济学家》(中国)，等等。

有关一带一路参考资源库的更多信息，请访问: https://www.ebscohost.com/academic/one-belt-one-road-reference-source。

(编译自: https://www.ebsco.com/news-center/press-releases/ebsco-information-services-helps-global-researchers-prepare-for-the-one-bel)

(本刊讯)